DOCUMENT RESUME

ED 328 275 IR 053 434

AUTHOR Li, Tze-chung
TITLE Information Retrieval in Multimedia Sources in an Electronic Age.
PUB DATE 88
NOTE 19p.
PUB TYPE Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Cataloging; *Databases; Gateway Systems; *Information Technology; *Microforms; *Online Searching; *Optical Data Disks; Standards; Telecommunications
IDENTIFIERS DIALOG; ORBIT

ABSTRACT

The last three decades have shown a marked development in new technologies for storing and retrieving information: microform in the 1960s; online database in the 1970s, and CD-ROM in the 1980s. While microform lacks volatility and multiple access points, preservation programs find it to be ideal storage media: it has a longer life expectancy than the other media; its use life is less limited; and it has a lower maintenance cost. Library operations have been significantly impacted by the large storage capacities and volatility of online databases and CD-ROMs. A few commercial publications provide information on additional physical forms in which a work is available, e.g., "Law Books and Serials in Print: A Multimedia Sourcebook" (Bowker, 1985) lists, in addition to books, audiocassettes, videocassettes, computer software, online databases, and microform. However, the shortage of bibliographical control of multisources, lack of data structure standardization, and the diversity of software programs often frustrate the user's information retrieval methods. Two charts compare and contrast the operation diversities of the BRS, Dialog, and ORBIT systems. Some of these diversities have been minimized through the emergence of gateway and the more sophisticated front end software, although a completely standardized interface which enables the users to access multiple systems has yet to be developed. (24 references) (MAB)

INFORMATION RETRIEVAL IN MULTIMEDIA SOURCES IN AN ELECTRONIC AGE

Tze-chung Li

ROSARY COLLEGE GRADUATE SCHOOL
OF LIBRARY AND INFORMATION SCIENCE
RIVER FOREST, ILLINOIS 60305, U. S. A.

1988

A cartoon that appeared in the Chicago Tribune depicts a reference librarian receiving a call. The party on the other line asks: "Is this the library? Please analyze Shakespeare's plays and give me a list of quotes reflecting aspects of existentialist thought. I'll hold on." It is humanly impossible for the reference librarian to provide this information in a short period of time. However, in an electronic age, it may not be a mission-impossible. A television series, the Knight Rider, dramatizes a car that not only can converse like a human being but is also able to provide answers to any questions asked. In a scenario envisioned by science fiction writers, a reader can browse through a computer catalog and call for books which will be delivered to him by a robot from the stacks, or a reader can retrieve three dimensional information from an encyclopedia or a hard copy dropped on his lap.[1] Though it sounds like a fantasy, the reality may not be too far away.

In his keynote speech at the 1987 EDUCOM conference, John Sculley mentioned three technologies that offers a new environment for gathering knowledge: hypermedia delivering information beyond the traditional media; simulation pushing the

knowledge boundary from static to dynamic; and finally, artificial intelligence.[2] Dialog Information Service added to its service in 1988 the provision of images in the TRADEMARKSCAN - FEDERAL database.[3] The Encyclopaedia Britannica will market very soon its Compton Electronic Encyclopedia in hypermedia in which the reader will be able to have texts, sound, and image in color.

In the last three decades, we have seen a significant array of new technologies for storing and retrieving information: microform in the 1960s, online database in the 1970's, and CD-ROM in the 1980's. Microform is a generic term including micro-print, microfilm, microcard, microform, and microfiche. The development of microform goes back to 1839 and microfilming printed products began in the 1930's.[4] In the 1960's, microforms became a fashionable phenomenon for information storage. In 1969 alone, the Library of Congress produced more than 11.3 million individual microfilm exposures on negative films and more than 5.6 million feet of positive print film from these and other negatives.[5] The density of microform recording ranges from 4 times to 1,000 times or even higher of its originals. A conventional microfiche provides about 900 pages per fiche. A microform with high range of density has the capacity of recording about 2,000 pages per fiche and can hold, with extreme range of density, the full text of a bible within one square inch of space. Microforms are excellent for storage and preservation and the least expensive. But the search of data in microform is slow and it lacks volatility and multiple access points.

Although the growth of microform has not been as fast as many people had anticipated, there has been an increased interest in recent years in using microform for storage and preservation of materials. Charles Chatwyck-Healey compared microform with videodisc technology and concluded that microforms will continue to be used until electronic technology in storage and transmission of data equals or improves microform in quality and cost.[6] Alan Calmes regards microfilm the best choice for storage and preservation on three grounds: its life expectancy is longer than all other media; its use life is far less limited; and its maintenance cost is lower.[7] He also reports that the Library of Congress continues to regard microform as a mainstay of its preservation programs with optical discs as a supplementary technology.[8] The publication of dual media, print and microform, such as those published by Congressional Information Service, is another example of continued interest in microform.

The introduction of online database in the 1970s has shaped the library operation in many ways. Access to materials is no longer physically limited to a library where materials are housed. It contains a large quantity of data yet provides fast retrieval capability. A computer search of a million records only takes a few seconds. The most useful feature of a database is its capability of providing multi-access points and a multi-dimensional approach. A searcher may locate items through author, title, source, language, date of publication, document type, and or any other access point imaginable. Users can also ask the computer to match the terms and retrieve only those documents in which a match occurs.[9] In this respect, it is far

superior to microform.

In the 1980s, the optical disc has caught the attention of libraries. An optical disc is a high density platter for writing or reading through a laser emitted beam. Data are impressed forming microscopic pits on the disc surface. There are four kinds: CD-ROM (compact disc - read only memory), CD-WORM (compact disc - write once read many times), CD-I (compact disc - interactive), and CD-Erasable (compact disc - erasable). CD-I holds digital text, images, sound animation, and graphs, a multi-me ia disc such as the Compton Electronic Encyclopedia just mentioned. By function, there are character encoded discs, video discs, and music discs. One side of a 5-inch disc can hold up to 72 minutes of digitability encoded music.[10]

A single CD-ROM can store 540 megabytes of data or hold 16,000 pages of data equivalent to approximately 1,500 floppy discs. In this capacity, the entire Encyclopaedia Britannica can be stored on one disc.[11] CD-ROM technology has gained rapidly in popularity. A study, CD-ROM Market Opportunities, by Link Resources, predicts that "between now and the end of the decade, CD-ROM players and information products will generate 2.3 billion in revenue for hardware, media, software, service, and information providers."[12]

Both online databases and CD-ROM expedite locating and disseminating information. In using these new technologies, the speed, storage capacity, and volatility in retrieving information are fascinating. Users are, however, also frustrated at the problems in locating these sources and the method of retrieving

information from them. They are specifically concerned about the shortage of bibliographical control of multisources, lack of standardization of data structures, and the diversity of software programs for information retrieval. We will take a closer look at these problems.

An ideal bibliographical control is to list complete records of human communication, to indicate location where records can be found, and to provide access to their contents. For online databases and CD-ROMs, there is no shortage of bibliographies of listing them separately. A few bibliographies of multisources have been published. Government Reports Announcement & Index, v. 75- , 1975- , lists titles in various forms: hardcopy, microform, and computer tape. The U. S. Bureau of the Census Catalog and Guide provides its products in the form of printed reports, microfiche, computer tapes, online access, diskettes, and maps. But the information is confined to its own publications.

Commercial publications have a more extensive coverage. Ulrich's International Periodical Directory, 1932- , has a section on serials available online. EBSCO's The Serials Directory: An International Reference Book, 1986- , provides information on additional physical forms, that is media other than a serial's original or conventional form. There are, however, many omissions. The bibliography of law books is perhaps the first such extensive bibliography of multisources. Law Books and Serials in Print: A Multimedia Sourcebook (1985, New York: Bowker, 6 volumes with quarterly cumulations), part of Bowker's Legal Reference System, lists, in addition to books,

audio cassettes, video cassettes, software, online databases, and microform. For more convenient use, there is a need for a bibliography that provides information on where to find materials, the available media in which the materials are contained, and the location of such media. It would be ideal for the National Union Catalog and the New Serials Titles to take the helm in providing multisources, listing not only print products but also other media available on a particular title.

In a library, the library catalog is a bibliographical control device that provides a record of materials that a library owns. Anglo-American Cataloguing Rules is designed as a guide to cataloging fourteen different kinds of materials, i.e. books, pamphlets, printed sheets, cartographic materials, manuscripts, music, sound recordings, motion pictures, videorecordings, graphic materials, computer files, three-dimensional artifacts and realia, microforms, and serials.[13] The computer file refers to "a file (data and/or program) encoded for manipulation by computer."[14] By this definition, CD-ROM and online databases are included. It is commendable that the Rules generally keeps path with new developments, yet, there is still room for refinement.

One obvious advantage of print products over computer files is that the former can be browsed. A reader opens a book, glances it over, and acquires immediately some information about its contents. A computer file is not visible to the naked eye and has a different degree of retrieval capability. It calls for more depth of description to present adequately its subject, content, coverage, update frequency, and device for retrieval.

Indication of its relationship to other media would be extremely useful.

It appears that the traditional concept of cataloging only for in-house materials and for object in hand remains the basic consideration of the Anglo-American Cataloging Rules. The procedure is not clearly outlined for cataloging online databases which are not physically in the library. The AACR is silent on the following aspects: (1) databases with different titles used by vendors; (2) databases with different sizes marked by vendors; (3) varied parallels, either broader or narrower, or lack of some features in the counterpart of a database or vice versa; and (4) databases with different abbreviations or file numbers for their titles.[15] Other problems include the amount of detail needed to describe the database and the number of added entries to be used. With reference to levels of detail in the description, Article 9.0D of the AACR refers to Article 1.0D under general rules for description, that is primarily for print products and is not adequate for online databases.

The effective bibliographical control of databases suggests the following:[16]

1. Title with medium designator in the bracket, parallel title, and other titles;

2. The parallel with other media, noting differences and range of coverage;

3. Search title or abbreviation or file number for access;

4. Period of coverage, updated and cumulated periodically;

5. Added entries for vendors that market the database;

6. Added entries for a series of databases;

In database structure, all files contain records and all records are divided by fields or paragraphs, and most fields are directly searchable. But arrangement of records and fields of a record vary with vendors. Take the American Statistics Index for example. When it was marketed at one time through ORBIT (On-line Retrieval of Bibliographic Information Timeshared), the Index sub-records were arranged under main records, thus each record contained many sub-records. The use of the logical operator AND in this file because of its record size (main records include many sub-records) will result in a large recall and is therefore not advisable. LINK and LINK NOT must be used in lieu of AND and AND NOT for more selective retrieval. In contrast, Dialog's arrangement of all sub-records as a single record results in smaller size of records. The use of AND and NOT will not retrieve as large a recall as ORBIT.

The ERIC file is available in BRS, Dialog, and ORBIT. ERIC provides some twenty-five searchable fields and several sub-fields. The file is reloaded by vendors into their own systems. Each vendor retains most of ERIC searchable fields, but deletes some sub-fields, and adds its own field, such as Dialog accession number. BRS does not use single word, major subject, and minor subject in the identifier field. It does not separate accession numbers for update, whereas ORBIT omits minor subject in both descriptor and identifier fields. It must be noted that the deletion of some sub-fields has little impact on information retrieval. Such a variation is usually ignored by users. However, the most frustrating for user may be the diversity in

labeling the fields by the various systems. Each field is generally labeled by two characters. There is no uniformity of labeling. In the searchable fields of ERIC, only five similar labels are used by BRS, Dialog, and ORBIT, and another five similar labels are used by two of the three systems; the remaining fields are labeled uniquely by each system.

Not all databases have the same fields and not all similar fields use same labels. In reviewing the fields of databases in three systems, the following data are found:[17]

	Number of Fields		
	BRS	Dialog	ORBIT
Basic Index		67	44
Additional Index		400	156
Total	208	467	200

It may be noted that searching is generally classified into basic index searching, or subject searching, and additional index searching, or non-subject searching (called subject-implicit searching in ORBIT). BRS does not make such a distinction. Not only do the number of fields and their availability in each file confuse the user, what is more confusing is that the same label can designate different fields, or the same field can be designated by different labels. For instance, in Dialog, the field CC= and N.Y.= designate more than ten different fields each.

Records in CD-ROM are similarly arranged, consisting of many fields, most of which are searchable. Bowker's Books in Print Plus has two search modes: search and browse. There are nineteen

fields in the search mode. The browse mode consists of nine fields. The Wilsondisc also provides two modes of searching. The browse mode consists of subject and personal name search. In the search mode, it provides six searchable fields: subject words, author, title, journal name, organization, and Dewey number. The search variations are far less than those available in Books in Print Plus. One unique feature of Wilsondisc is that it consists of more than one file of the Wilson family.

ERIC CD-ROM is currently marketed through three vendors. Although the data structure is designed by ERIC, variations in retrieval exist. The SilverPlatter's ERIC CD-ROM provides some twenty-nine searchable fields and sub-fields, whereas Dialog ERIC Ondisc's fields for menu search is quite limited. The ease of use, however, does not lie in the number of fields, but in the search mechanism.

Practically all searching in CD-ROMs for information retrieval is structured in menus. The user must search information step by step according to the displayed menu instructions. It is time-consuming, yet easy to grasp with little or no training. But the degree of ease varies. In Books in Print Plus, nineteen search fields are displayed in the search mode, each is represented by two characters followed by an equal sign "=". The user selects a field and inputs information for the system to search. The user may select another field and combine the search results. Although searching process is simply and straight-forward, there are caveats to keep in mind. First, a title search must be given exact words. If the title contains

the word, AND, OR, or AND NOT, the entire title must be in quotations so that the system will ignore these words as logical operators. Second, there are inconsistencies in input of the word, such as "the" and "and." The "&" sign or the word "and" are used inconsistently. The search will fail if the user keys in the title with the word "and" in it, when the system uses "&" instead of "and," or vice versa. Another inconsistency arises with the word "the" which is included in some cases, but omitted in others. The search for the title "The Garden Jungle", for example, should be entered as "Garden Jungle" or it will result in zero posting. For title search, if exact words are not known for certain, using the browse mode is advised.

SilverPlatter is the most convenient to use. Although basically a menu search, it imparts features of free-text search. The user may enter any terms and phrases and disregard the stop words. The user may, for instance, enter "pursuit of happiness," "gone with the wind," or "government of the people." The system will retrieve documents in which these words are present in the same field. Hyphenated phrases are searched as descriptors. Other features include display of index terms and downloading.

Dialog's Ondisc provides both menu and command searches. Its command search is the same as most Dialog online database commands. Its menu search is designed for a step by step search. The user may limit the search result by adding another term or using other options. It is also capable of downloading. Though it lacks the menu volatility of SilverPlatter, its Dialog command search is superb and sets off the limited menu field search well.

The concept of searching is basically the same for all

these systems. These systems all enable users to search any word in the text, match the searched word, search the words in a string or near one another, limit the search to type, year, date, or language of publication, combine terms with author or other fields, and repeat the same search strategy in different files. But in operation details, each system differs from the other in access, protocols, commands, and output format. Operation diversities may be grouped as follows: (1) different procedures for access; (2) different commands with the same result; (3) different features with compatible results; and (4) features not common to all systems. The chart below lists some features not common to all systems:

Features (Y/N)	Systems BRS	Dialog	ORBIT
Security Code in addition to password	Y	N	Y
Online thesaurus	N	Y	N
SELECT STEPS	N	Y	N
SHOWSELECT (display of all terms in the SELECT list	N	N	Y
Negative Qualifier	Y	N	N
LINK/LINK NOT	N	Y	N
SELECT after each term EXPANDed/ROOTed required	Y	Y	N
Online sorting	Y	Y	N
Searched terms highlighted	Y	N	N
Print field in any sequence	N	N	Y

Images	N	Y	N

A benefit from diversity is the competition for excellence. But in many aspects, it serves no meaningful purpose. Simone Klugman has questioned the validity of having different commands for logoff, varied symbols for truncation, and inconsistent formatting of citation.[18] In addition, one may wonder why logon, stacking command, and field labels cannot be standardarized. The lack of uniformity and standardization not only among the systems but also within the system itself is an area most users consider an inconvenience. The user must memorize different protocols, commands, field labels, and other different features in switching from one system to another.

Over the years, the three systems have made noticeable progress toward compatibility, though not uniformity, in logon, search, result, and logoff procedures. Below are some of the achievements made by the three systems:

1. The Stacking command (/) in BRS can be changed so as to conform with the other two systems.
2. The ..SET DETAIL=ON command in BRS and the AUDIT=ON command in ORBIT achieve similar results as Dialog's Super Select.
3. The BREAK function in BRS is no longer limited to stopping print.
4. The introduction of proximity searching is a great feature by ORBIT to achieve compatibility with other systems.
5. BRS has made available its online sorting capability.
6. Tailored print command is now available in Dialog.
7. ORBIT uses LOGOFF as an alternative to its STOP.
8. In displaying dictionary terms, EXPAND can be used as an alternative to NEIGHBOR in ORBIT.

In online retrieval, one significant development in easing diversity is the gateway facility by which one system serves as a revolving door to other system. The gateway enables one system to be linked to other system and/or other databases,[19] such as the Westlaw to Dialog, ALANET to EasyNet, and OCLC to BRS, just to name a few. An obvious advantage of such a gateway is that the user needs one logon procedure saving the steps of providing communication protocols, identification number, and passwords to different systems.

The emergence of gateway and front end software for microcomputer use is another milestone towards minimizing the problems stemming from of diversity. The gateway is defined as the type of software that takes the user "to the entrance of the databank (the gate) but no farther."[20] Of the many software, Perfect Link, Cross Talk, and Smartcom are three well-known packages.[21] Smartcom, for instance, is able to perform an automated logon function.

Front end software is more sophiscated. Over a dozen front end software are available on the market, including DialogLink, Pro-Search, Sci-Mate, and Wilsearch. The front end software features, in addition to gateway functions, access to several systems, pro-search editing and uploading, database selection, and post-processing of search results.[22] Since all front end packages use menu driven searches, one drawback is time-consumption, and its repeated step-by-step queries most tedious.

Diversity and lack of standardization in CD-ROM remain a big problem to tackle. It would be an ideal to develop a program that makes the use of different CD-ROMs compatible. In online

searching, efforts have been made towards standardization. In 1980, Subcommittee Z 39G (Standard Terms, Abbreviations, and symbols for Use in Interactive Information Retrieval) of the American National Standard Committee X-39 was formed for the purpose of reducing existing diversity in the command languages presently in use. At the same time ISO (International Organization for Standardization) Technical Committee 46/Sub-Committee 4/Working Group 5 was formed for the same purpose. Standardization takes many years to achieve and requires close cooperation, extensive discussion, continued debate, and eventual compromise. Complete standardization may never be accomplished. What we need now is an interface that enables us to use one system to retrieve information on the other.

It is reported that Lawrence Livermore National Laboratory (LLNL) has developed an intelligent gateway that allows a user to access diverse database resources. The software handles all communication, logon, and access procedures.[22] In 1986, an agreement was reached by participants of the Linked System Project (LSP). In this project, the four participants, the Library of Congress (LC), the Research Libraries Group (RLG), the Western Library Network (WLN), and On-Line Computer Library Center (OCLC) will set standards that permit each of them link to other for searching.[23] The latest service offered by H. W. Wilson is for its database Licensing Service to provide machine readable magnetic tapes of databases for local online access. The service is adaptable to NOTIS (Northwestern Online Total Integrated System) commands and protocols. Both ORBIT and BRS

were purchased by the same company. It has been reported that an interface will be developed in about six months to enable the use of BRS commands to search in ORBIT or vice versa.

Evidently, the problem of diversity is already becoming less of a problem. It would not be at all surprising to see a spurt of developments by 1995 for convenient, effective retrieving information in multimedia sources. Bibliographies of multimedia sources of information will emerge in different forms. The user will be able to locate multimedia sources from one source. There will be increased use of subject-based rather than medium-based materials regardless of their forms. A user may use a single medium that contains information in whatever form, such as text, films, slides, photographs, and drawings. There will be intelligent programs by which a user may automatically logon, select database and systems, retrieve information, display the result, and logoff. An interface in making diverse systems compatible with each other to access information will become a reality. The interface will be an integrated system that has the capabilities also for word-processing, spell-checking, spreadsheet and graphic presentations, and database management. The user may single out the information retrieved with automatic citation; edit, re-format and re-process it; input the piece of information into his own datafile for future use; and present, on the basis of retrieved information, statistical data or graphic illustrations. It would not be beyond belief to see the full development of artificial intelligence application of which can select systems, begin particular databases, retrieve relevant information, analyze the information retrieved and reach a

decision, and even generate new concept and idea.

REFERENCES

1. Colin Steele, "From Punched Cards to Robots: Our Ascent into Technology," Wilson Library Bulletin, 62(2): 39-32, 31 (1987).

2. Keynote speech at 1987 EDUCOM, OCLA videotape #17562069.

3. Chronolog 16 (1): 1 (1988).

4. Walt Crawford, Current Technologies in the Library: An Informal Overview. Boston: G. K. Hall, 1988, pp. 26-27.

5. Charles G. La Hood, Jr. "Microfilm for the Library of Congress," College and Research Libraries, 34 (1973): 291-94.

6. Charles Chadwyck-Healey, "The Future of Microform in an Electronic Age," Wilson Library Bulletin 58 (4): 270-3 (1983).

7. Alan Calmes, "New Confidence in Microfilm," Library Journal, 111(15): 38-42 (1986).

8. Ibid., 40.

9. Tze-chung Li, An Introduction to Online Searching (Westport, CT: Greenwood Press, 1985) , pp. 8-9.

10. William R. Nugent,"Optical Discs - An Emerging Technology for Libraries," IFLA Journal, 12 (3): 175-81 (1986).

11. Henry Fersko-Weiss, "Data Storage: How Much Do You Need? Personal Computer 9(6): 64-71 (1985).

12. Information Today, 3 (6): 22 (1986).

13. J ed., rev. (Chicago: American Library Association, 1988, 677p.)

14. Ibid., p. 617.

15. Li, op. cit., p. 71.

16. Ibid., pp. 71-2.

17. Ibid., pp. 264-7.

18. Simone Klugman, "Online Information Retrieval Interface with Traditional Reference Services," Online Review, 4: 271 (1980).

19. Ruth N. Cuadra and Carlos A. Cuadra, "In Search of Gateways," Information Today 4(3): 7 (1987).

20. Donald T. Hawkins and Louise Levy, Front End Software for Online Database Searching, Part I: Definitions, Systems Features, and Evaluation," Online 9(6): 31 (1985). In her earlier writing, Levy defined gateway broadly to include what we now call front end. See Louise Levy, "Gateway Software: Is It for You?" Online 8 (6): 67-79 (1984).

21. For a comparison of Perfect Link and Crosstalk, see Maurita Peterson Holland, "Communications Software: Experience with Perfect Link and Crosstalk XVI," Online 8(4): 75-80 (1984).

22. Hilary D. Burton, "Technology to Provide Excellence in Information Services," Special Libraries 78(1): 1-6 (1987).

23. Richard W. McCoy, "The Linked Systems Project: Progress, Promise. Realities," Library Journal 111: 33-39 (1986).

24. Coined after the terms used by Pat Molholt. see Molholt, "A View from the Chip: The Influence of Information Technologies on Libraries and Librarianship," INFLA Journal 13 (10): 16 (1987).

END

U.S. Dept. of Education

Office of Educational
Research and Improvement (OERI)

ERIC

Date Filmed
July 24, 1991